Comparing LancsBox and AntConc in the extraction of Passives and Nominals: Towards Objectivity in Critical Discourse Analysis

H. Zih, M. El Biadi, and Z. Chatri

***Abstract*—Natural language processing programs geared for the analysis of large linguistic corpus have exponentially proliferated thanks to machine-readable texts. Critical discourse analysts have eventually become more interested in adopting corpus linguistic approaches to discourse analysis. The present paper aims to compare two widely renowned corpus linguistics programs, namely LancsBox and AntConc to analyze linguistic features viz. passive and nominalized constructions. The study seeks to evaluate the software effectiveness in culling passive and nominalized structures from a large scale of linguistic data. We will after that proceed to comparing the software findings to the results obtained through a manual analysis of the data to see if there are any differences as well as the extent to which critical discourse analysis combined with corpus linguistic methods can offer more objective and reliable results, refuting by this the common cited criticism of attempting to prove a preconceived point. A corpus of ten news articles published in *The Times* online newspaper were downloaded and analyzed both manually and digitally so as to examine the occurrences and the distribution of the two aforementioned grammatical constructions in the news reports. The findings show that corpus linguistic software can reliably extract passive and active instances from the texts. Although both LancsBox and AntConc revealed approximately the same frequencies compared to the findings we obtained manually, both programs did not help in specifically isolating the clauses which report passivized and nominalized actions performed by the perpetrator. Therefore, the study concluded that while the corpus linguistic software can facilitate the identification of frequent and salient linguistic patterns especially in a large scale of data, the human interpretation is mandatory as far as the research purpose is concerned.**

***Index Terms*—AntConc, Corpus linguistics Software, Critical Discourse Analysis, Grammatical Structure, LancsBox, Nominalization, Passivisation.**[1]

I. INTRODUCTION

The present study seeks to compare two well-known software in corpus linguistics used to analyze linguistic data, namely LancsBox and AntConc. The main objective is to assess the extent to which these two programs can accurately and effectively identify two linguistic structures viz. passivisation and nominalization in a set of ten news reports downloaded from *The Times* online newspaper. Linguists adopting the Critical Discourse Analysis approach (henceforth CDA) have often been criticized for being subjective and biased when dealing with different linguistic forms and the way they are used in discourse, for they mostly rely on the qualitative and not the quantitative method to analyze their data, and the offered interpretations are not based on systematic language description. [1],[2]. McEnery argues that people who do CDA often have a large corpus to choose from, but since they have to take so many aspects into account, they tend to 'cherry pick' their data so that 'a really detailed in-depth qualitative analysis' can be carried out [3]. This raises the issue that the selected text is often the result of the critical discourse analyst own biases. For instance, if the linguists using CDA think the press reports issues related to Muslims negatively, they are more likely to select a small number of texts that prove that. Thus, the findings might confirm that those 'cherry picked' texts are really 'Islamophobic', but the question that [4] raises is: 'How about the 6000 texts which might represent Muslims and Islam positively?'. Therefore, the present paper aims to evaluate the extent to which corpus linguistics software can assist critical discourse analysts identify the linguistic features they are interested in allowing a more manageable and objective analysis to be undertaken. Therefore, to counter the criticism that critical discourse analysts tend to prove a 'preconceived point', a corpus-based approach will be adopted. Using natural language processing programs can assist the researcher to process large amounts of linguistic

[1] This paper was submitted for review on April 25th, 2021

Hanane Zih is a PhD researcher at the Faculty of Letters and Human Sciences Dhar Mehraz, Sidi Mohamed Ben Abdullah University, Fez, Morocco (e-mail: hanane.zih@usmba.ac.ma)

Maha El Biadi is a professor at the Faculty of Letters and Human Sciences Dhar Mehraz, Sidi Mohamed Ben Abdullah University, Fez, Morocco (e-mail: maha.elbiadi@usmba.ac.ma)

Zakariae Chatri is a PhD researcher at the Faculty of Sciences Dhar Mehraz, Sidi Mohamed Ben Abdullah University, Fez, Morocco (e-mail: zakariyae.chatri@outlook.fr)

data, making it possible for them to avoid being subjective to a certain degree while handling different types of texts. This is achieved by enabling them to come up with quantitative analyses that can point to some discernible patterns, which they can then interpret and explain by linking them to their broader social context of use.

II. CORPUS LINGUISTICS SOFTWARE

Researchers such as [5],[6], [7], [8], [9] have attempted to integrate general methodological approaches of corpus linguistics with CDA arguing that “corpus approaches help to reduce researcher bias” [2, p.8]. In the same vein, Tankard [10] describes the researcher criteria, while approaching a text, as ‘vague’ and ‘subjective’ suggesting that this shortcoming can be addressed through ‘a systematic identification of linguistic elements and structural dimensions” which can be made possible thanks to computer-assisted analysis [10]. When a researcher decides to adopt a corpus linguistic approach, the used procedures are not biased as computers are not biased when analyzing texts. McEnery [5] asserts that “[a computer] does not pick out certain things because it thinks they are interesting or they confirm its own suspicions. […] they kind of direct us to things that we maybe wouldn’t have thought of ourselves. And that is really good, because it stops us from being biased.” [5].

Computer-assisted text analysis using tools such as LancsBox and AntConc can be of great help to discourse analysts whose objective is to study large corpora of data with the aim of drawing some good generalizations concerning the way certain linguistic forms are used in discourse. Not only can such software help highlight the main patterns in particular corpus, but they can also direct the researcher’s attention to ‘minority discourses’ which are stated less often. Analyzing the data manually, though precise and accurate in identifying the forms the researcher is interested in, it can only be used to process a restricted amount of data, and thus the chances of missing such ‘minority patterns’ are higher [5]. This constitutes a limitation, which can undermine the validity of the conclusions that the researcher can draw. Not only does corpus linguistic software enhance validity and reliability, but also efficiency of research [11] by saving the analyst a substantial amount of time and effort. Thanks to the computer’s superior capacity in processing a large corpus in no time, researchers can efficiently test their hypotheses against large collections of data which would have been, otherwise difficult to achieve.

Researchers using corpus linguistics software often distinguish between two different types of corpus approaches viz. corpus-driven approach and corpus-based approach. The former is “an inductive process where corpora are investigated from the bottom up and patterns found therein are used to explain linguistic regularities and exceptions of the language variety/genre exemplified by those corpora.” [5]. This approach is based on viewing the data from ‘a naïve perspective’. McEnery [5] points out that though it is not viable to be entirely naïve when analyzing data, the analysts have to start off with that idea in mind and try not to impose any views or ideas or hypothesis. The corpus should drive the analysis based on the obtained frequencies and patterns. The second approach, on the other hand, is a more traditional CDA perspective as the critical discourse analysts begin their analysis with hypotheses which have been formulated based on a certain way of representing a group. It is ‘where corpora are used to test performed hypotheses or exemplify existing linguistic theories.” [5]. With the hypotheses in mind, the researcher uses a corpus to investigate certain linguistic items so as to see whether there are any emerging patterns.

McEnery [5] argues that combining the two approaches is ‘the best kind of analysis’ since, according to him, carrying out a corpus analysis from a very naïve approach is “really difficult […] we always have ideas about what a certain group have been represented in, because we live in society, and we are aware of how they are being talked about in advance.” Similarly, in a study conducted by Baker *et al.* [7] investigating discourses of immigration in 140-million-word corpus of British newspaper texts, they argued that the adopted ‘recursive model’ allows ‘moving back and forth between quantitative and qualitative forms of analysis, with each stage informing the subsequent stage’ [7, p.248]. Hence, combining CDA with corpus linguistics [12], [13], [7], [14] has proved to yield more robust and valid set of findings since the researchers’ interpretation is thus ‘grounded in systematic language description’ [1, p.148].

The following section outlines the corpus linguistics software to rely upon in the current study, namely LancsBox and AntConc. In addition to the aforementioned benefits of combining corpus linguistics with CDA, LancsBox and AntConc were selected for application particularly because of the up-to-date functionalities they offer, which render the analysis process quite straightforward. Also, both software are free to access and easy to use compared with other similar software packages.

A. LANCSBOX

LancsBox is a new-generation software package for the analysis of language data and corpora developed at Lancaster University [15]. The free software can be used by linguists, language teachers, historians and anyone interested in language as it allows the researchers to work either with their own data or the existing data. LancsBox offers powerful searches at different levels of corpus annotation. It comes with a number of features including *KWIC tool* (Key Word In Context) which generates a list of all instances of a search term in a corpus in the form of a concordance, *Whelk tool* which provides information about how the search term is distributed across corpus files and *Words tool* that allows in-depth analysis of frequencies of types, lemmas and POS categories as well as comparison of corpora using the keywords techniques [4].

B. ANTCONC

AntConc is a freeware, multi-platform, multi-purpose corpus analysis toolkit, designed by Laurence Anthony in Waseda university in Japan [16] for teaching and learning purposes. It was first released in 2002 to assist students with their technical writing course at the Osaka University Graduate School of Engineering. In December 2004, [16] released an improved version of the software, AntConc 3.0. It includes an extensive set of text analysis tools viz. KWIC Concordance, Search Term Distribution Plot, Original File View, Word Clusters/ Lexical bundles, Word lists, and key word lists among others. It also has powerful search features such as Regular Expressions (REGEX) and Extensive Wildcards which allow for a more complex analysis. The concordance tool is the key feature of AntConc software because “a concordance program can find and display a huge number of examples in varied contexts and situations quickly and efficiently.” [16].

III. Benefits of Computer-Assisted Text Analysis

Computer-assisted text analysis can be considered potentially ‘more objective’ compared with manual text analysis. Murphy points out that text analysis software is free from ‘the presuppositions knowingly or unknowingly imposed by the researcher’[11, p.284] because the text analyst does not have to abide by any procedures, such pre-reading, coding and pre-specifying categories or concepts to be identified in the text. Being independent from the researcher’s presuppositions, the software generated findings are more objective and valid, as a result.

Another significant advantage of computer-assisted text analysis is research reliability. As long as there is a fixed algorithm in the computer software, multiple researchers can replicate the process with the same texts resulting in the same findings. The reliability of digital text analysis is 100% [17]. For example, if another researcher is interested in using a particular text analysis software to examine the frequencies of passives and actives in the same texts deployed in the present study, the obtained percentages should be exactly the same.

Furthermore, computer-assisted text analysis is a cost-effective approach, for computers are capable of processing a large corpus which cannot be manually handled in much shorter timeframes, enhancing by this research efficiency [17]. Considering the merits of computer-assisted text analysis while undertaking multidisciplinary investigations, the present study is based on corpus linguistics software, LancsBox and AntConc, in order to see the extent to which the manually obtained results are effective and valid enhancing by this objectivity in critical discourse analysis.

IV. Critical Discourse Analysis

It is useful to briefly define the concept of critical discourse analysis because it does feed in quite a bit to the present study. CDA views language as a social practice, and is mainly concerned with the way ideologies and power relationships are expressed through language and in a text. Van Dijk [18] defines it as

> *A type of discourse analytical research that primarily studies the way social power abuse, dominance, and inequality are enacted, reproduced, and resisted by text and talk in the social and political context. With such dissident research, critical discourse analysts take explicit position, and thus want to understand, expose, and ultimately resist social inequality.* [18, p.352].

The work of critical discourse analysts starts from the premise that the grammatical structure is not devoid of meaning. It aims, therefore, to uncover the hidden meanings underlying the linguistic structure in a given piece of discourse.[19],[20],[21],[22],[23],[24],[25],[26],[27].

CDA is relatively complex as it functions at different levels, which makes it quite challenging, and complicated to sometimes do. Critical discourse analysts do take into account issues of production and reception as it seeks to investigate the text and power relations in, between, and behind the constructions aiming by this to offer an in-depth analysis of texts. Having to describe the texts and discursive patterns and relationships and links them to social situations and contexts, critical discourse analysts often choose to analyze a limited number of texts to be able to offer a thorough analysis. Richardson [28] points out that CDA is a bridge that connects both society and its sociopolitical issues with the critical analysis of language. Therefore, we argue that corpus linguistics can assist critical discourse analysts not only analyze large corpora, but it is believed to diminish the potential ambiguities and make it possible to critical discourse analysts to confidently draw conclusions and generalizations which are grounded on systemic language description.

V. Research Design

As mentioned above, this study seeks to draw some comparisons between two computer-assisted text analysis programs, LancsBox and AntConc, to test their ability to identify the occurrences of passive and active structures and nominalized constructions in ten news reports collected from *The Times* online newspaper. It will after that compare their findings to those obtained through a manual analysis to see if the findings are similar. It might be argued that the number of texts is not sufficient to reach valid conclusions pertaining the effectiveness of the two software under investigation. In fact, while the corpus linguistics software can handle large data, researchers can only analyze a limited number of texts thoroughly, for analyzing a larger data sample manually would be laborious and time-consuming. For the same reasons, the researchers opt for collecting their own data instead of working on pre-uploaded texts into the software.

The research questions guiding this comparative study are the following:
- Are there any similarities and /or differences between the manually obtained frequencies and those generated digitally?
- If so, what are these similarities and /or differences?
- What factors did contribute to these similarities and/or differences?
- To what extent can corpus linguistics software enhance objectivity in CDA?

The present study has a threefold objective:

- To compare the findings obtained from LancsBox and AntConc with those obtained manually.
- To evaluate the effectiveness of corpus linguistics software in extracting passive and nominalized constructions.
- To determine the extent to which computer-assisted text analysis can enhance objectivity in CDA.

To achieve our aim, we divided the collected data into two sets of news reports, each of which talks about two different categories of protagonists. The first set of articles contains five news stories reporting criminal actions carried out by individuals which, for research purposes will be called group A. The second set of articles comprises articles talking about criminal acts perpetrated by individuals, which we will categorize as group B. Going into further detail relating to the identity of the individuals that the news reports talk about is of little relevance for the present study, and doing so will only overshadow the main objective of our investigation. It is however worth mentioning that these two groups have different characteristics, which distinguish them from each other. The focus, at present, is to find out whether the automated analysis of these two sets will display a statistical difference as far as passive and nominalized constructions are concerned. It will on the one hand compare the two computer programs to see if they give similar results and then compare the results of the manual analysis of the data with the results of the computer programs to evaluate the extent to which the two programs are accurate and reliable in identifying the linguistic forms under study. Examining the extent to which the manually obtained findings can be objective is the ultimate aim of the present study.

An important number of studies has been done adopting both CDA and corpus linguistics on different texts to show their hidden agendas or specific themes. The present paper is different as it is the first one to compare the effectiveness of LancsBox and AntConc in extracting passive and nominalized constructions. The findings of this study provide a critical evaluation of these linguistics software and highlights the factors explaining the potential similarities and differences between the manual results and the digital ones. It will also be helpful for researchers interested in adopting corpus linguistics to get a glimpse of the functionalities of LancsBox and AntConc which are believed to be time saving and efficient especially when working on a large amount of data.

The diagram below shows the algorithm implemented in the context of this study.

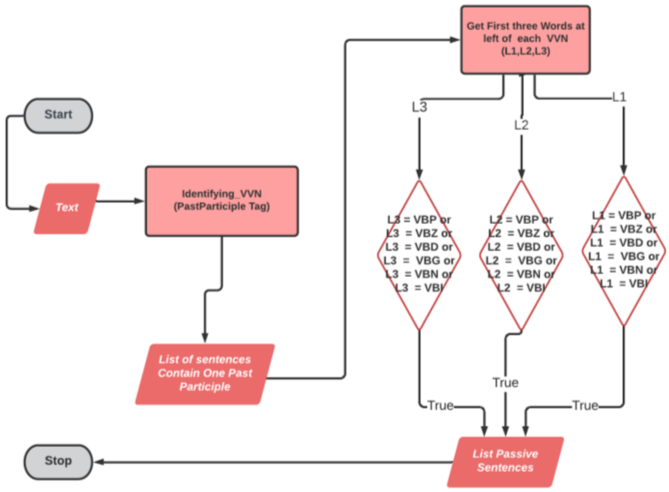

Get First three Words at left of each VVN (L1,L2,L3)
Start
L3
L2
L1
Text
Identifying_VVN (PastParticiple Tag)
L3 = VBP or L3 = VBZ or L3 = VBD or L3 = VBG or L3 = VBN or L3 = VBI
L2 = VBP or L2 = VBZ or L2 = VBD or L2 = VBG or L2 = VBN or L2 = VBI
L1 = VBP or L1 = VBZ or L1 = VBD or L1 = VBG or L1 = VBN or L1 = VBI
List of sentences Contain One Past Participle
True
True
True
Stop
List Passive Sentences

Fig. 1. The implemented Algorithm

The two grammatical features that we are concerned with, as mentioned earlier, are passive and nominalized constructions. As to passivized constructions, they provide 'a convenient way of postponing the agentive subject by turning it into the agent in a passive construction' [29, p. 416]. The following are examples to illustrate:

1- The book was written by John. (A passive construction where 'John' the doer of the action is mentioned last in the clause).
2- John wrote the book. (John is the agentive subject in an active construction. It occurs initially in the clause)

As to nominalized constructions, they come about as a result of the syntactic process whereby a verb is converted into a noun [30],[31]. Example 3 below contains the nominalized form 'criticism', which is derived from the verb 'to criticize', which is used in example 4:

3- These ideas have been subject to widespread criticism.

4- Many people have criticized these ideas. [31, p.168].

In this study, the researchers have adopted a corpus-based approach in the analysis of news reports as, it has been mentioned earlier, the current paper is based on ongoing research; therefore, the study has been carried out with the research hypotheses in mind. Nevertheless, the corpus-driven approach cannot be undermined as the automated findings drive the researchers to go back and forth before making any final conclusions. First of all, the researchers, after a close reading of the two sets of news articles, manually identified and calculated the occurrences of passivisation and nominalization in each corpus. After that, the corpora were loaded and imported into LancsBox and AntConc for analysis as it is shown in the following graphs:

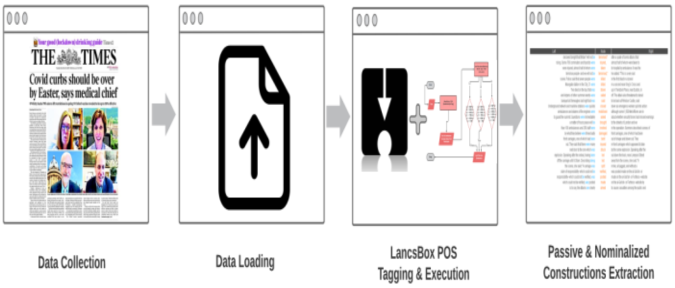

Data Collection
Data Loading
LancsBox POS Tagging & Execution
Passive & Nominalized Constructions Extraction

Fig. 2. LancsBox Pipeline

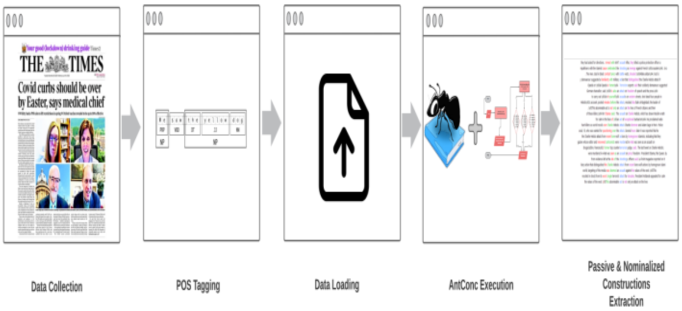
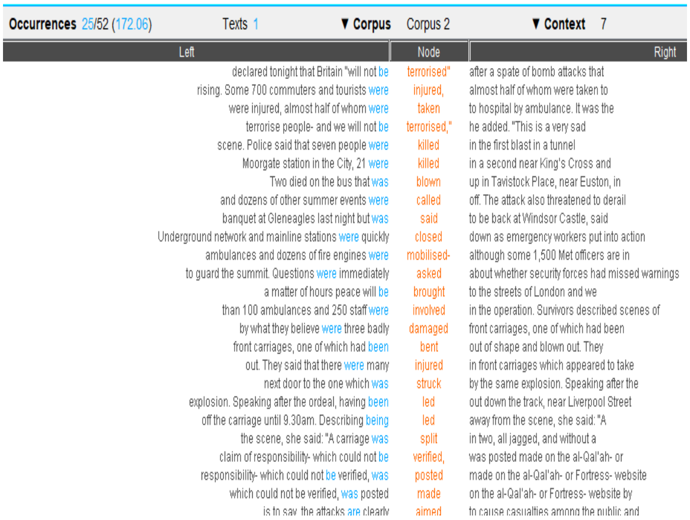


Fig.3. AntConc Pipeline

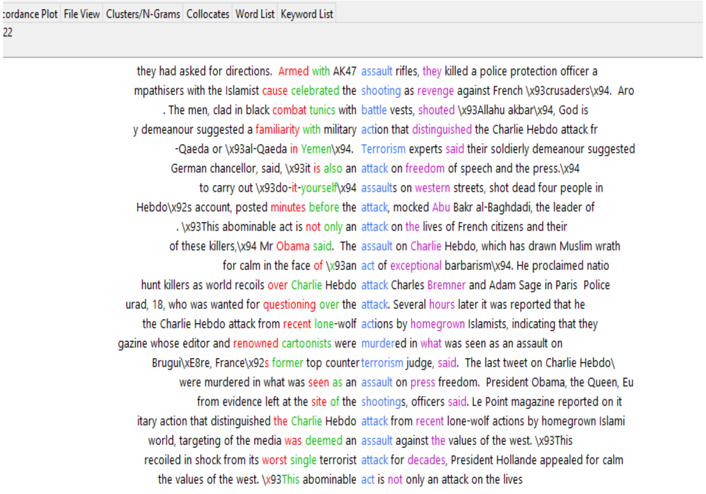
While the loading was quite simple with LancsBox as it can read different formats, AntConc supports only '.txt' and '.pdf' formats so the corpora were transformed into '.txt' format before loading. To generate a list of passives and nominals in the corpus, the KWIC tool specifically 'the advanced search' tool, was used while working with LancsBox. KWIC displays basic information about the frequency of the search term and its distributions in texts. By means of Parts of Speech Tagging we could extract all the words with the tag 'VVN' to identify the past participle and then use the setting 'context' which changes the number of words that are displayed in the concordance, which allows users to look at words in context, to the left and to the right of the 'node' (search term) as it is shown in the following figure.


Occurrences 25/52 (172.06) Texts 1 ▼ Corpus Corpus 2 ▼ Context 7

Left	Node	Right
declared tonight that Britain "will not be	terrorised"	after a spate of bomb attacks that
rising. Some 700 commuters and tourists were	injured,	almost half of whom were taken to
were injured, almost half of whom were	taken	to hospital by ambulance. It was the
terrorise people- and we will not be	terrorised,"	he added. "This is a very sad
scene. Police said that seven people were	killed	in the first blast in a tunnel
Moorgate station in the City, 21 were	killed	in a second near King's Cross and
Two died on the bus that was	blown	up in Tavistock Place, near Euston, in
and dozens of other summer events were	called	off. The attack also threatened to derail
banquet at Gleneagles last night but was	said	to be back at Windsor Castle, said
Underground network and mainline stations were quickly	closed	down as emergency workers put into action
ambulances and dozens of fire engines were	mobilised-	although some 1,500 Met officers are in
to guard the summit. Questions were immediately	asked	about whether security forces had missed warnings
a matter of hours peace will be	brought	to the streets of London and we
than 100 ambulances and 250 staff were	involved	in the operation. Survivors described scenes of
by what they believe were three badly	damaged	front carriages, one of which had been
front carriages, one of which had been	bent	out of shape and blown out. They
out. They said that there were many	injured	in front carriages which appeared to take
next door to the one which was	struck	by the same explosion. Speaking after the
explosion. Speaking after the ordeal, having been	led	out down the track, near Liverpool Street
off the carriage until 9.30am. Describing being	led	away from the scene, she said: "A
the scene, she said: "A carriage was	split	in two, all jagged, and without a
claim of responsibility- which could not be	verified,	was posted made on the al-Qal'ah- or
responsibility- which could not be verified, was	posted	made on the al-Qal'ah- or Fortress- website
which could not be verified, was posted	made	on the al-Qal'ah- or Fortress- website by
is to say, the attacks are clearly	aimed	to cause casualties among the public and

Fig. 4. KWIC Sample

On the other hand, the corpus linguistic technique 'concordance' was used while working with AntConc. It should be mentioned that in order to display the instances of our search term, AntConc cannot create its own model of tags, so loading our tagged data was mandatory. The following figure illustrates this further:



```
            they had asked for directions.  Armed with AK47 assault rifles, they killed a police protection officer a
 mpathisers with the Islamist cause celebrated the shooting as revenge against French \x93crusaders\x94.  Aro
          . The men, clad in black combat tunics with battle vests, shouted \x93Allahu akbar\x94, God is
  y demeanour suggested a familiarity with military action that distinguished the Charlie Hebdo attack fr
             -Qaeda or \x93al-Qaeda in Yemen\x94.  Terrorism experts said their soldierly demeanour suggested
            German chancellor, said, \x93it is also an attack on freedom of speech and the press.\x94
                  to carry out \x93do-it-yourself\x94 assaults on western streets, shot dead four people in
      Hebdo\x92s account, posted minutes before the attack, mocked Abu Bakr al-Baghdadi, the leader of
            . \x93This abominable act is not only an attack on the lives of French citizens and their
            of these killers,\x94 Mr Obama said.  The assault on Charlie Hebdo, which has drawn Muslim wrath
                 for calm in the face of \x93an act of exceptional barbarism\x94. He proclaimed natio
      hunt killers as world recoils over Charlie Hebdo attack Charles Bremner and Adam Sage in Paris  Police
    urad, 18, who was wanted for questioning over the attack. Several hours later it was reported that he
      the Charlie Hebdo attack from recent lone-wolf actions by homegrown Islamists, indicating that they
  gazine whose editor and renowned cartoonists were murdered in what was seen as an assault on
       Brugui\xE8re, France\x92s former top counterterrorism judge, said.  The last tweet on Charlie Hebdo\
           were murdered in what was seen as an assault on press freedom.  President Obama, the Queen, Eu
               from evidence left at the site of the shootings, officers said. Le Point magazine reported on it
   itary action that distinguished the Charlie Hebdo attack from recent lone-wolf actions by homegrown Islami
       world, targeting of the media was deemed an assault against the values of the west. \x93This
     recoiled in shock from its worst single terrorist attack for decades, President Hollande appealed for calm
          the values of the west. \x93This abominable act is not only an attack on the lives
```

Fig. 5. Concordance Sample

VI. FINDINGS

A. *Passivisation in the Corpora*

The results revealed that the manually obtained frequencies of passives show that passive constructions are more frequent in the set of articles about group A compared to group B of news reports. AntConc and LancsBox results relatively confirmed this finding.

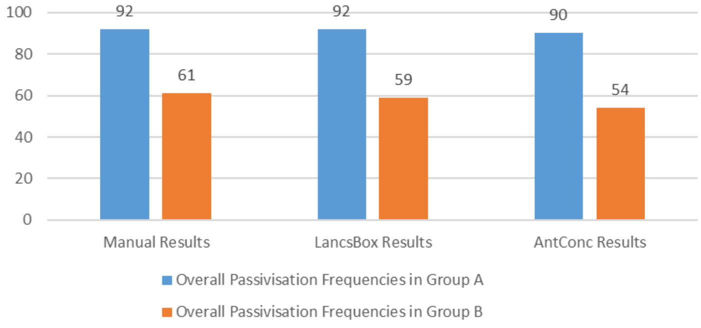


Fig. 6. Overall Passivisation Frequencies in the Corpora

A close reading of the above graph reveals that LancsBox passive frequencies are the same compared to the ones carried out manually (92 occurrences) of passives in group A as the graph above shows while AntConc missed two instances of passives. However, the total number of passive constructions that was obtained manually is 61 instances in group B compared to 59 with LancsBox and 54 with AntConc.

B. *Nominalization in the Corpora*

Since the researchers' main concern is to identify the nominalized forms that directly refer to the actions performed by the protagonists that the two groups of newspaper texts talk about, we carefully analyzed the texts to identify all the instances of nominals, such as ‘shooting’, ‘attack’. The main criterion is to consider only nouns that are derived from verbs and are used to refer to the acts that were performed either by individuals belonging to group A or those belonging to group B. The quantitative analysis revealed that the frequency of nominalized constructions in the group A was higher compared to that of group B. This is mainly due to the fact that group A contains lengthier sentences compared to group B, and by consequence, a higher number of clauses with more nominalized structures. (Group A contains 629 clauses, whereas group B comprises 351 clauses in total).

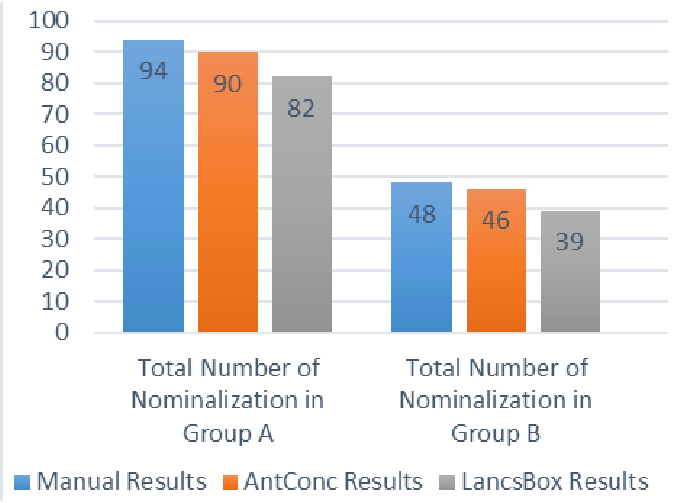


Fig. 7. Overall Frequencies of Nominalization Frequencies in the Corpora

Although there are some small differences in the number of identified occurrences of nominalizations, it is clear that both AntConc and LancsBox computer assisted text analysis software could both reveal that the number of nominals is higher in article set A than in set B. While LancsBox, could identify all the coded nominals in the corpora, AntConc not only could generate the coded nominals, but also instances containing what can be seen as a verb form embedded within the noun. For example, instead of extracting only the noun ’act’, the program also highlights nouns such as ‘***act***ions’ and ‘f***act***ory’ as they contain the three letters that make up the noun ‘act’. This accounts for the fact that the number of features that AntConc could identify is higher compared to that generated by LancsBox. Making some alterations in AntConc to avoid the extraction of inaccurate instances is not possible without the ‘source code’ of the software. The question that arises however is whether all the detected instances using AntConc are actually nominalizations in the grammatical sense of the word. Do the extracted nouns all derive from verbs? The noun ‘factory’ mentioned above is a case in point; a critical discourse analyst will certainly not consider it a nominalized process derived from the verb to act (See example 3 and 4 above). It is etymologically unrelated to it. That said, when we compare the results of AntConc to those of the manual analysis, we can clearly see that they are the closest to the analysis carried out manually compared to LancsBox.

VII. DISCUSSION

Overall, we can at this point confidently say that both programs under investigation have managed to identify in a systematic and successful way the passives and the nominals in the corpora. Neither program, however, was successful at isolating instances of passive constructions according to the doer of the action that the process of the clause talks about. In order for that to be achieved, it is inevitable for the analyst to resort to the manual method, at least for the time being, for a more fine-tuned analysis that also takes into consideration the semantic meaning of language and not just its syntactic structure.

It should be made clear at this point that the manual results are roughly the same with the automated ones. The small number of passives and nominalized instances which were missed by AntConc and LancsBox, as the graphs above show, can by no means prove that the manually obtained findings are less objective at least as far as the current study is concerned. Although the researchers started off by a corpus-based approach, it does not imply that they have certain presuppositions to confirm. The found differences are mainly due to the research questions as we aim to investigate the number of passives and nominalized constructions of the actions performed only by the perpetuator. Therefore, it is of paramount importance to question the efficiency of such computerized tools as the LancsBox software developer, Brezina, clearly declares that "the tools themselves restrict us. They only allow us to do certain things. They do not allow us to do everything." [4]. And one obvious thing that neither LancsBox nor AntConc could identify is the passive linguistic structures which refer only to the perpetuators, and nominalizations which are derived from verbs and directly refer to the agents. Critical discourse analysts who are willing to rely on machine-readable text to analyze large scale of data and thus come up with more objective, valid and reliable findings will face such kind of limitations. [4] goes even further to claim that the tools themselves are subject to biases as the people who design such tools are biased themselves. He said "they […] incorporated their own biases and interests into the way that they have created those tools. Today's tools may lead us down certain paths, or put us in certain mindsets, or ways of thinking." [4]. Interestingly, the computer-assisted text tools can themselves be biased which leads one to question the extent to which they can yield objective results; therefore, it is of paramount importance to resort to the researcher's interpretation to objectively address the research questions.

Since CDA is mainly concerned with investigating "the way social power abuse, dominance, and inequality are enacted, reproduced and resisted by text and talk in the social and political context." [18, p. 352]. The human interpretation can by no means be superfluous, for contextual factors, such as culture and media genre should be taken into account in identifying and analysing the linguistic structures. While computer-based approaches can assist critical discourse analysts in examining a large corpus, which would otherwise be time consuming and laborious if done manually, such contextual factors are impossible to be taken into consideration while performing the analysis automatically. The prominence of the software is inevitable in identifying frequencies of passives and nominals, yet the human interference is always needed 'to identify the larger tale and the broader schemas' in which these linguistic structures are woven. Combining the quantitative and the qualitative approaches along with empirical and interpretative examination would make of the software's value more explicit if compared with the analytical approach to the manual analysis [32].

VIII. CONCLUSION

In this study, we sought to compare AntConc and LancsBox with the aim of evaluating their effectiveness in extracting passives and nominal forms from a corpus of news media texts. The main concern was to explore the extent to which computer-assisted text analysis software can assist researchers in the area of discourse analysis to reach objective and reliable results. Even though isolating the clauses that directly refer to the perpetuator was not possible using AntConc and LancsBox, the findings show that both AntConc and LancsBox are efficient overall in the identification of roughly all the passive constructions and nominals.

The employment of this kind of corpus linguistics software to process linguistic data is believed to offer an answer to the problem of subjectivity that CDA is criticized for. It makes it possible for the researchers to obtain some quantitative results that can display certain meaningful patterns, helping them, thus, reach a certain degree of objectivity in their analyses of discourse. The use of these methods can also facilitate their task by limiting the data down to manageable amounts, and by assisting them to extract only the parts that specifically contain the features they are interested in investigating. It will enable them to save time and effort in their endeavor to study large amounts of data, and will by this, help them to overcome the reluctance that they generally feel towards engaging in this type of quantitative studies.

REFERENCES

[1] H.G. Widdowson, "The theory and practice of critical discourse analysis," *Appl. Lings*, vol.19, no.1, pp. 136–151, 1998.

[2] P. Baker, "Acceptable bias? Using corpus linguistics methods with critical discourse analysis," CDS, vol.9, no.3, pp.247-256, 2012, doi: 10.1080/17405904.2012.688297.

[3] V. Koller and G. Mautner, "Computer application in critical discourse analysis,", in *Applying English grammar: Corpus and Functional approaches*, *C.* Coffin, A. Hewings and K. O'Halloran, Eds. London: Arnold, pp.216-228.

[4] Lancaster University Online. (2021). Corpus linguistics: Method, analysis, interpretation. [MOOC]. Available: https://www.futurelearn.com/courses/corpus-linguistics/8/todo/84759

[5] G. Hardt-Mautner, "Only Connect: Critical Discourse Analysis and Corpus Linguistics", UCREL Technical Paper 6, Mar.2021. [Online]. Available: http://ucrel.lancs.ac.uk/tech_papers.html

[6] M. Stubbs, *Text and Corpus Analysis*, Oxford: Blackwell, 1996.

[7] P. Baker, C. Gabrielatos, M. Khosravinik, M. Krzyzanowski, T. McEnery, and R. Wodak, "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press," Discourse and Society, vol. 19, no. 3, pp. 273– 306, 2008, doi: 10.1177/0957926508088962

[8] J. Morley and P. Bayley, Eds. *Corpus Assisted Discourse Studies on the Iraq Conflict: wording the war*. London: Routledge, 2009.

[9] A. Partington, "Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: an overview of the project," Corpora, vol. 5, no.2, pp. 83- 108, 2010.

[10] J. W. Tankard, "The empirical approach to the study of media framing," in *Framing Public Life*, S. D. Reese, O. H. Gandy, and A. E. Grant, Eds., Mahwah, NJ: Lawrence Erlbaum Associates, 2001, pp. 95–106

[11] P. Murphy, "Affiliation bias and expert disagreement in framing the nicotine addiction debate," Since., Tech., H. V., vol.26, no. 3, pp. 278-299, 2001.

[12] R. Krishnamurthy, "Ethnic, racial and tribal: The language of racism?," in *Texts and practices: Readings in critical discourse analysis*, C.R. Caldas-Coulthard and M. Coulthard, Eds., London: Routledge, 1996, pp. 129 –149.

[13] M. Hoey, "A clause-relational analysis of selected dictionary entries: Contrast and compatibility in the definitions of 'man' and 'woman'," in *Texts and practices: Readings in critical discourse analysis*, C.R. Caldas-Coulthard and M. Coulthard, Eds., London: Routledge, 1996, pp. 150 –165.

[14] P. Baker and T. McEnery, "The value of revisiting and extending previous studies: The case of Islam in the UK press, in *Quantifying Approaches to*

Discourse for Social Scientists: Post disciplinary Studies in Discourse, R. Scholz, Ed., Palgrave Macmillan, Cham, 2019, doi:10.1007-3-319-97370-8_8

[15] *LancsBox v.5.x*. (2020). V. Brezina, P. Weill-Tessier and A. McEnery. Accessed: Dec. 12, 2020. [Online]. Available: http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_6.0_manual.pdf

[16] *AntConc v.3.5.8*. (2019). L. Anthony. Accessed: Dec. 20, 2019. [Online]. Available: http://www.laurenceanthony.net/software/antconc/

[17] G. Shapiro, "The future of coders: Human judgments in a world of sophisticated software," in *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, C. W. Roberts, Ed., Mahwah, NJ: Lawrence Erlbaum Associates, 1997, pp. 225-238.

[18] T.A. Van Dijk, "Critical discourse analysis," In *Handbook of discourse analysis*, D. Tannen, D. Schiffrin and H. Hamilton, Eds., Oxford: Blackwell,2001, pp. 352-371.

.[19] N. Fairclough, *Analysing Discourse. Textual analysis for social research*, UK: Routledge.

[20] N. Fairclough, J. Muldrerrig, and R. Wodak, "Critical Discourse Analysis". In *Discourse as Social Interaction,* T.A. Van Dijk, Ed., London: Sage, 2011, pp.357-378.

[21] R. Fowler, *Language in the News: Discourse and Ideology in the Press.* London: Routledge and Kegan Paul, 1991.

[22] T.A. Van Dijk, "Discourse as Interaction in Society," in *Discourse as Social Interaction*, vol,2., T. A. van Dijk, Ed., London: Sage, 1997, pp.1-37.

[23] Van Dijk, T.A. "Critical Discourse Studies." In *The Handbook of Discourse Analysis*, D. Tannen, H. E. Hamilton and D. Schiffrin, Eds., Oxford: OUP (in press), 2012.

[24] G. Weiss and R. Wodak, "Introduction: Theory, Interdisciplinarity and Critical Discourse Analysis," In *Critical Discourse Analysis: Theory and interdisciplinarity*, G. Weiss and R. Wodak, Eds., London: Palgrave Macmillan, 2002, pp. 1-32.

[25] R. Wodak, "The Discourse Historical Approach," In *Methods of Critical Discourse Analysis*, R. Wodak and M. Meyer, Eds., London: Sage, 2001, pp.81-115.

[26] R. Wodak, "Complex Texts. Analyzing, Understanding, Explaining and Interpreting Meanings," *Discourse Studies*, vol. 13, no.5, pp. 623-633, 2011a.

[27] R. Wodak, "Critical discourse analysis: Challenges and perspectives," In *Critical Discourse Analysis: Volume 1 Concepts, History, Theory*, R. Wodak, Ed., London: Sage Publications Ltd, 2013, pp. 22-45.

[28] J.E. Richardson, *(Mis)Representing Islam: The racism and rhetoric of British broadsheet newspapers*. Amsterdam: John Benjamins, 2004.

[29] S. Greenbaum and R. Quirk, *Student's Grammar of the English Language*, Longman, 1990.

[30] N. Fairclough, *Analysing Discourse: textual analysis for social research*, New York: Routledge,2003.

[31] G. Thompson, *Introducing Functional Grammar*, London: Arnold, 1996.

[32]M. Touri and N. Koteyko, " Using corpus linguistic software in the extraction of news frames: towards a dynamic process of frame analysis in journalistic texts," International Journal of Social Research Methodology, Vol. 18, no.6, pp. 601-616, 2015, doi: 10.1080/13645579.2014.929878